Explainable AI: A Review of Machine Learning Interpretability Methods

Year: 2020 | Citations: 2277 | Authors: Pantelis Linardatos, Vasilis Papastefanopoulos, S. Kotsiantis

---

Abstract

Recent advances in artificial intelligence (AI) have led to its widespread industrial adoption, with machine learning systems demonstrating superhuman performance in a significant number of tasks. However, this surge in performance, has often been achieved through increased model complexity, turning such systems into “black box” approaches and causing uncertainty regarding the way they operate and, ultimately, the way that they come to decisions. This ambiguity has made it problematic for machine learning systems to be adopted in sensitive yet critical domains, where their value could be immense, such as healthcare. As a result, scientific interest in the field of Explainable Artificial Intelligence (XAI), a field that is concerned with the development of new methods that explain and interpret machine learning models, has been tremendously reignited over recent years. This study focuses on machine learning interpretability methods; more specifically, a literature review and taxonomy of these methods are presented, as well as links to their programming implementations, in the hope that this survey would serve as a reference point for both theorists and practitioners.